Bioinformatics, 2018, 1–8
doi: 10.1093/bioinformatics/bty842
Advance Access Publication Date: 5 October 2018
Original Paper

Sequence analysis

DeepHINT: understanding HIV-1 integration via deep learning with attention

Hailin Hu[1], An Xiao[2], Sai Zhang[3], Yangyang Li[4], Xuanling Shi[4], Tao Jiang[5,6,7], Linqi Zhang[4], Lei Zhang[1,*] and Jianyang Zeng[2,*]

[1]School of Medicine and [2]Institute for Interdisciplinary Information Sciences, Tsinghua University, Beijing 100084, China, [3]Department of Genetics, Stanford Center for Genomics and Personalized Medicine, Stanford University School of Medicine, Stanford, CA 94305, USA, [4]Comprehensive AIDS Research Center, Collaborative Innovation Center for Diagnosis and Treatment of Infectious Diseases, School of Life Sciences and School of Medicine, Tsinghua University, Beijing 100084, China, [5]Department of Computer Science and Engineering, University of California, Riverside, CA 92521, USA, [6]Bioinformatics Division, BNRIST/Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China and [7]Institute of Integrative Genome Biology, University of California, Riverside, CA 92521, USA

*To whom correspondence should be addressed.
Associate Editor: Bonnie Berger

Received on April 4, 2018; revised on September 7, 2018; editorial decision on September 27, 2018; accepted on October 4, 2018

Abstract

Motivation: Human immunodeficiency virus type 1 (HIV-1) genome integration is closely related to clinical latency and viral rebound. In addition to human DNA sequences that directly interact with the integration machinery, the selection of HIV integration sites has also been shown to depend on the heterogeneous genomic context around a large region, which greatly hinders the prediction and mechanistic studies of HIV integration.

Results: We have developed an attention-based deep learning framework, named DeepHINT, to simultaneously provide accurate prediction of HIV integration sites and mechanistic explanations of the detected sites. Extensive tests on a high-density HIV integration site dataset showed that DeepHINT can outperform conventional modeling strategies by automatically learning the genomic context of HIV integration from primary DNA sequence alone or together with epigenetic information. Systematic analyses on diverse known factors of HIV integration further validated the biological relevance of the prediction results. More importantly, in-depth analyses of the attention values output by DeepHINT revealed intriguing mechanistic implications in the selection of HIV integration sites, including potential roles of several DNA-binding proteins. These results established DeepHINT as an effective and explainable deep learning framework for the prediction and mechanistic study of HIV integration.

Availability and implementation: DeepHINT is available as an open-source software and can be downloaded from https://github.com/nonnerdling/DeepHINT.

Contact: lzhang20@mail.tsinghua.edu.cn or zengjy321@tsinghua.edu.cn

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Integration of the HIV-1 genome to human genome is a crucial step in viral infection and replication cycle. Clinically, the integration of HIV is closely related to the formation of latent viral reservoir and the rebound of viral load when antiretroviral therapy (ART) is interrupted (Wong *et al.*, 1997). Furthermore, recent studies have also

revealed that the integration of HIV provirus within specific genes can affect the persistence of infected cells (Maldarelli *et al.*, 2014; Wagner *et al.*, 2014), indicating a more significant role of the selection of HIV integration sites in disease progression.

Despite the long-lasting research efforts, the detailed mechanisms and functional implications of the selection of HIV integration sites still remains largely unclear (Lusic and Siliciano, 2017). In addition to the local sequence motifs of the human genome that directly interact with the DNA integrase (Serrao *et al.*, 2014), previous researches have also associated the preference of HIV integration events with various genomic landmarks, e.g. the binding of integrase cofactor LEDGF/p75 (Ciuffi *et al.*, 2005), actively transcribed genes (Schröder *et al.*, 2002), intron regions (Singh *et al.*, 2015), chromatin accessibility (Vijaya *et al.*, 1986) and nuclear landscape (Marini *et al.*, 2015). To integrate diverse genomic features for predicting HIV integration sites, several computational methods have been proposed (Berry *et al.*, 2006; Santoni *et al.*, 2010). However, these methods strongly rely on explicit feature engineering and input from various experimental data, e.g. RNA-seq, ChIP-seq and DNase-seq data, which may not be universally available for all the integration prediction tasks. In addition, the sequence resolution and the scope of feature engineering also limit the interpretation of mechanistic insights from these methods, leading to the insufficient usage of currently available large-scale HIV integration data (Shao *et al.*, 2016).

Nowadays, in computational biology, deep learning has become the state-of-the-art prediction methods in many applications, e.g. identification of nucleotide-protein binding sites (Alipanahi *et al.*, 2015; Zhang *et al.*, 2015), prediction of the functional effects of noncoding sequence variants (Quang and Xie, 2016; Zhou and Troyanskaya, 2015) and translation process modeling (Zhang *et al.*, 2017a,b). On the other hand, despite the superior prediction performance, the explainability and the understanding of feature organizations of deep learning models often lag behind, which not only limits the applicability of deep learning techniques in exploring unknown cellular mechanisms and gaining insights, but also raises potential concerns of using a black box. One possible strategy to increase the explainability of deep learning models is the introduction of attention mechanisms, which are particularly designed to extract important regions of input data by training an additional neural network that learns the relative importance of each input position from local features (Bahdanau *et al.*, 2014). Thus, when applied to analyze the genomic sequence data, the introduction of attention mechanisms is expected to reveal important sequence positions that shape the prediction results from the deep learning framework and thus provide potentially important mechanistic insights about the observed genomic phenomena (Deming *et al.*, 2016; Mao *et al.*, 2017; Pan and Yan, 2017; Singh *et al.*, 2017).

Here, we have developed an attention-based deep learning framework, named DeepHINT (Deep learning for HIV INtegraTion) (Fig. 1), for accurately predicting HIV integration sites by automatically extracting important features and genomic positions from primary DNA sequences alone or together with epigenetic information. The validation and analysis results showed that the DeepHINT model with DNA sequence as input alone (denoted by DeepHINT seq) can possess sufficient prediction power and provide important biological implication, demonstrating the learning ability of deep learning in extracting useful sequence features from the context of HIV integration sites. In addition, DeepHINT is flexible to incorporate other types of genomic data, such as H3K36me3 ChIP-seq, (denoted by DeepHINT seq+H3K36me3), which further boosted the prediction power and facilitated a better identification of attainable positions from the genome context. Our work

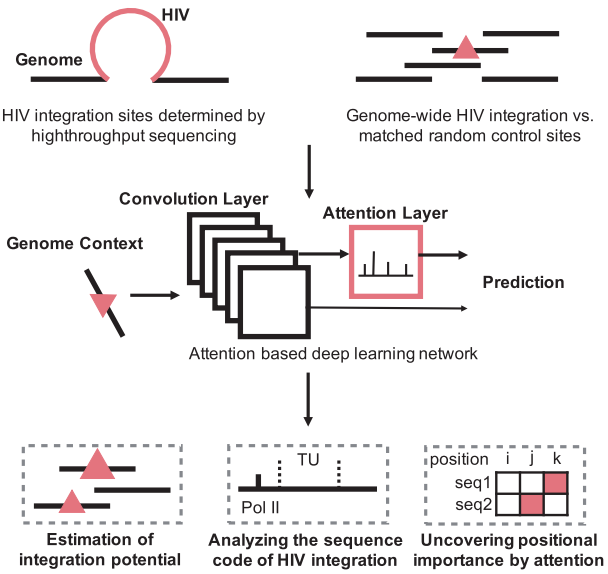

Fig. 1. Schematic overview of the DeepHINT pipeline. In our prediction task, we used experimentally derived HIV integration sites as positive samples. To account for potential bias caused by enzyme digestion in the sequencing process, matched random control sites possessing the same distance distribution to the nearest enzyme digestion sites were generated as negative samples. See the main text for more details

represents the *first* attempt to model the selection of HIV integration sites by deep learning approaches. Extensive tests have shown that DeepHINT can achieve an accurate prediction performance and outperform the current state-of-the-art prediction methods that leverages way more experimental data. The biological relevance of the DeepHINT prediction results has also been validated by the associations with known genomic markers of HIV integration. More importantly, the information derived from the incorporated attention mechanism clearly indicates the relative importance of each position in the input genomic context of the predicted HIV integration sites, which can thus help us to explain both local and distal genetic features captured by the deep learning model. In particular, our prediction results highlight the potential roles of several DNA-binding proteins, which may expand our current understanding of HIV integration site selection. All these results have demonstrated the effectiveness of our deep learning based prediction approach and also provided useful insights to facilitate the mechanistic studies of the HIV integration process.

2 Materials and methods

In this section, we will describe the details of our deep learning model for HIV integration site prediction. For simplicity, in the text we mainly focus on the DeepHINT model with sequence information as input alone. We also provide a stepwise mathematical description of the model architecture and the training process in the Supplementary Notes to facilitate a better understanding of our deep learning model.

2.1 Feature extraction by a convolutional neural network

DeepHINT first employs multiple convolution-pooling modules to automatically learn informative sequence features in the surrounding sequences of HIV integration sites (Fig. 2a). In particular, we

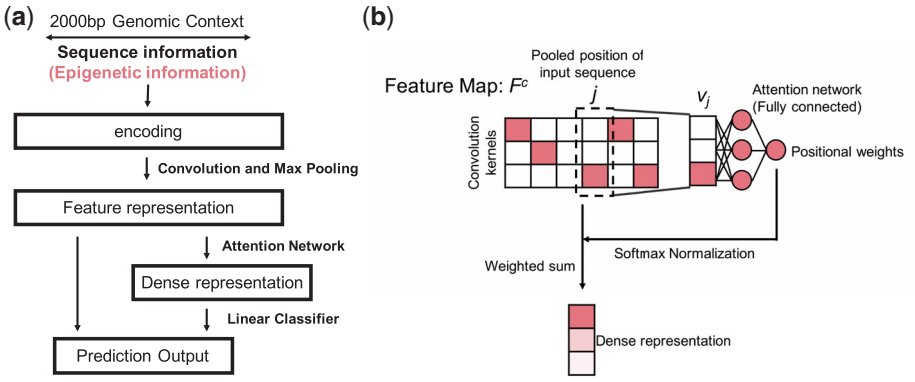

Fig. 2. The deep learning framework implemented in DeepHINT. (**a**) The overall schematic view of the deep learning framework. (**b**) The illustration of attention mechanism. See the main text for more details

first extend each HIV integration site both upstream and downstream by 1000 bps to obtain the sequence context, resulting in a sequence profile denoted by $s = (nt_1, ..., nt_{2000})$, where nt_i stands for the nucleotide at the ith position. Each nucleotide in the sequence profile is then converted to a binary vector of length 4 by one-hot encoding, with each dimension corresponding to a nucleotide type. In the convolutional layer, a series of one-dimensional convolution operations are performed over the 4-channel input data, in which each channel corresponds to one dimension of the binary vector. In particular, each convolution operation corresponds to a weight matrix (i.e. kernel) that can also be regarded as a position weight matrix (PWM).

More specifically, given a genomic sequence $s = (nt_1, ..., nt_{2000})$ and the corresponding one-hot encoded representation E, the convolutional layer computes $X = \text{conv}(E)$, i.e.

$$X_{k,i} = \sum_{j=0}^{p-1} \sum_{l=1}^{4} W_{k,j,l} E_{l,i+j}, \tag{1}$$

where $1 \le i \le 2000 - p + 1$, $1 \le k \le d$, p is the kernel size, d is the kernel number and W is the kernel weights. Next, we apply the rectified linear activation function (ReLU) on the convolution results, which mimics the biological neuron activation. After convolution and rectification, the max-pooling operators are used to perform dimension reduction. Therefore, through a series of convolution-pooling modules, the sequence profile can be compiled to a $d \times q$ feature map matrix (denoted by F^c), where q represents the total (pooled) positions of the input sequence (Fig. 2b).

2.2 Incorporation of the attention mechanism

To better capture and understand the positional importance of the sequence context, we further introduce an attention layer into our model (Fig. 2a). The attention layer takes the feature vector after convolution-pooling operations as input, and then computes a score indicating whether the neural network shall pay attention to the sequence features at that position. Basically, column j of the feature map matrix F^c can be viewed as a feature vector (denoted by v_j) that describes the features of the jth position in the input sequence, with each dimension corresponding to a kernel in the convolutional layer. The attention layer feeds each input feature to a shared feedforward neural network with a single hidden layer. The output of the attention layer is an importance score, denoted by e_j, for which a larger value indicates that the corresponding position is more important for the contribution to final HIV integration site prediction. In particular, the columns of the feature map matrix F^c are further averaged by taking the normalized importance scores α_j as weights, resulting in a dense feature representation F^a, i.e.

$$F^a = \sum_{j=1}^{q} \alpha_j v_j, \tag{2}$$

$$\alpha_j = \frac{\exp(e_j)}{\sum_{t=1}^{q} \exp(e_t)}, \tag{3}$$

where e_j is the importance score output by the shared neural network and α_j is the corresponding normalized score.

To integrate the features captured by the convolution-pooling modules (i.e. F^c) and the attention mechanism (i.e, F^a), we first concatenate all the values in matrix F^c and linearly project them to one value (denoted by S^c) that represents the contribution from a unified representation of the whole sequence. Finally, we concatenate S^c with the dense representation F^a and then feed them together to a logistic regression classifier to obtain a prediction score that indicates the probability of HIV integration. In summary, the full model can be expressed as

$$Pred(s) = \text{sigm}(\text{concat}(F^a, S^c)), \tag{4}$$

in which s denotes the genomic context of a candidate integration site and

$$S^c = \text{dense}(\text{pool}(\text{conv}(\text{encode}(s)))), \tag{5}$$

where encode(·), conv(·), pool(·), concat(·), dense(·) and sigm(·) represent the one-hot encoding, convolution, max pooling, concatenation, dense and sigmoid operations, respectively. Meanwhile, given a specific input sequence, we can also output a weight vector (denoted as *AttMap*)

$$AttMap(s) = (\alpha_1, ..., \alpha_q), \tag{6}$$

which expresses the model's attention on each position of the input sequence.

2.3 Model training

After hyperparameter calibration (Supplementary Notes and Supplementary Table S1), the deep neural network of DeepHINT is trained by minimizing the binary cross-entropy loss function, which is defined as the sum of negative log likelihood, i.e.

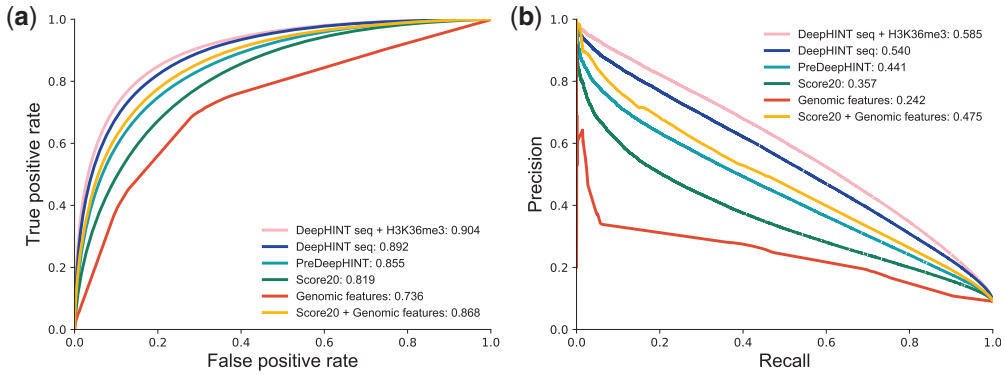

Fig. 3. Prediction performance on the test dataset. (**a–b**) Comparison of prediction performance of DeepHINT with that of different baseline methods, in terms of (a) receiver-operating characteristic (ROC) curves and (b) precision recall (PR) curves, respectively. 'DeepHINT seq' denotes the DeepHINT framework with DNA sequence alone as input, while 'DeepHINT seq +H3K36me3' denotes the DeepHINT framework with DNA sequence as well as H3K36me3 ChIP-seq data as input. 'preDeepHINT' denotes the DeepHINT seq framework without using the attention mechanism. The 'Genomic features' mean the genomic profiles collected from the ENCODE project and ChIP-Atlas (see Supplementary Notes)

$$loss = -\sum_i \log\left(y_i Pred(s_i) + (1 - y_i)(1 - Pred(s_i))\right), \quad (7)$$

where y_i stands for the true binary label of an input sequence s_i. The standard error backpropagation algorithm (Rumelhart *et al.*, 1986) and the batch gradient descent method (Bengio, 2012) are implemented for training. We also introduce several regularization techniques, including adding max-norm constraints on kernel weights (Srebro *et al.*, 2005), dropout (Srivastava *et al.*, 2014) and early stopping (Bengio, 2012), to alleviate the potential overfitting problem. In addition, to better address the data imbalance problem (i.e. the number of negative samples is much larger than that of positive samples) and make full use of the excessive negative sample information, we also apply an bootstrapping-based training strategy (Wallace *et al.*, 2011; Zhang *et al.*, 2017b). Specifically, in parallel, we train the deep neural network for 16 times with an equal number of randomly sampled (with replacement) positive and negative samples from the original training set. This strategy results in an ensemble of deep neural network classifiers, whose prediction scores and attention maps are then averaged to give the final output, i.e.

$$DeepHINTScore(s) = \frac{1}{16}\sum_{i=1}^{16} Pred_i(s), \quad (8)$$

$$AttMap(s) = \frac{1}{16}\sum_{i=1}^{16} AttMap_i(s), \quad (9)$$

where $Pred_i(s)$ and $AttMap_i(s)$ represent the prediction score and the output attention map of a single deep neural network for a given input sequence s, respectively.

The attention-based deep neural network of DeepHINT has been implemented using the Keras library 2.0.8 (https://keras.io), and one GTX 1080Ti GPU has been used to accelerate the training and testing processes. Using such a hardware setting, the training and testing process for each model of DeepHINT takes 34 min and 140 s using our training and testing set, respectively.

3 Results

3.1 DeepHINT accurately predicts HIV integration sites

We have performed extensive tests on known HIV integration sites in the HEK293T cell line (Singh *et al.*, 2015) obtained from the Retrovirus Integration Database (Shao *et al.*, 2016) and found that DeepHINT can significantly outperform the other state-of-the-art models in predicting HIV integration sites. As the experimental determination of HIV integration sites involved a restriction enzyme (*MSEI*) digestion step that may lead to bias in the sequence context, matched random control sites were generated as negative samples following the same protocol described previously (Berry *et al.*, 2006; Santoni *et al.*, 2010; Singh *et al.*, 2015; Wang *et al.*, 2007). More specifically, we first determined the genomic distances between all the positive samples to their nearest *MSEI* sites and then randomly sampled nine times more matched control sites that had the same distance distribution to their nearest *MSEI* sites as the negative samples. Note that the number of negative samples was set to be ten times as many as positive samples to reflect the natural imbalance of integration versus non-integration sites. To facilitate the training and evaluation process of our model, the whole dataset was separated into strictly non-overlapping training and testing sets by chromosomes. Specifically, samples on chromosomes 1, 2, 3 were assigned to the test set, while the samples from the remaining chromosomes were used as the training set. Overall, the aforementioned protocol resulted in 743 465 and 214 019 positive samples for training and test sets, respectively, as well as a corresponding ten times larger set of negative samples. The final prediction performance was evaluated and reported based on the test data.

We first compared our method with a conventional position weight matrix (PWM) based method, namely Score20 (Berry *et al.*, 2006) (Supplementary Notes), which directly calculates the consistency of the −10 to +10 bp window of a given site of interest with the consensus motif generated from training data. That is, Score20 mainly focuses on the local sequence motifs that favor HIV integrase binding and the window size has been confirmed to generate a satisfactory choice (Supplementary Fig. S1). Expectedly, by efficiently integrating a much broader genome context of HIV integration, DeepHINT achieved a great improvement over Score20, with an increase of the area under the precision recall (AUPR) curve by 18.3% and the area under the receiver-operating characteristic (AUROC) curve by 7.3% (Fig. 3a and b).

On the other hand, as also shown in the previous studies (Berry *et al.*, 2006; Santoni *et al.*, 2010; Singh *et al.*, 2015), the surrounding genomic features, e.g. chromatin accessibility, histone markers, transcription unit and intron annotation, also possess certain predictive information for detecting the retrovirus integration sites. Therefore,

to test whether DeepHINT can sufficiently capture the genome context of HIV integration sites, we also compared its prediction performance to that of a random forest based model which explicitly required these additional surrounding genomic features as input, both with and without incorporating the Score20 values representing the local DNA sequence features (Supplementary Notes). In particular, the following experimentally measured genomic profiles were used as input to this random forest based model, including the ChIP-seq data of H3K27Ac, H3K36me3, H3K4me1, H3K4me3, H3K9me3, RNA polymerase (Pol) II and CTCF, DNase-seq data, as well as transcription unit and intron unit labels derived from RNA-seq data (Supplementary Notes and Supplementary Table S2). We found that the random forest based model solely built on the above genomic features performed poorly (Fig. 3a and b), with an AUPR score of 24.2% and an AUROC score of 73.6%, indicating a necessity to effectively integrate various genomic context information with the Score20 values in the modeling process. Intriguingly, although the integration of these additional genomic data did boost the prediction performance of Score20, DeepHINT still outperformed the random forest model by 6.5% in AUPR and 2.4% in AUROC (Fig. 3a and b). Note that such a comparison was biased to the random forest model as DeepHINT only took DNA sequence as input while the random forest model was fed with plenty of additional experimental data that have been shown to correlate with HIV integration.

To confirm the above results, we also implemented two other machine learning models, i.e. logistic regression and gradient boosting decision tree (GBDT), which showed similar performance with random forest (Supplementary Table S3). Also, to exclude the possibility of introducing noise by combining multiple features (Santoni, 2013), we also compared the prediction of individual genomic features and confirmed the superior performance of our model (Supplementary Table S4). All these results demonstrated that the deep learning framework employed in DeepHINT can effectively learn the hidden feature representations encoded in the primary DNA sequences surrounding the HIV integration sites. Notably, we found that without using the attention mechanism, the prediction performance dropped significantly, with a decrease of 9.9% in AUPR and 3.7% in AUROC, which validated the contribution of attention mechanism to final prediction results of DeepHINT (Fig. 3a and b).

While the DNA sequence alone already generated good prediction performance, we further attempted to incorporate the cell type specific information in our framework. Here, we chose to use H3K36me3 as the information source as it has been shown to be the best single predictor among all the epigenetic profiles in our analysis (Supplementary Table S4). Expectedly, the incorporation of H3K36me3 further improved the prediction performance of DeepHINT, especially in terms of the precision recall curve, achieving an AUPR score of 58.5% and an AUROC score of 90.4% (Fig. 3).

To alleviate potential concerns about using a large number of low-frequency integration sites with substantial overlapping sequence context, we also constructed an additional high frequency integration dataset and tested the performance of different methods (Supplementary Notes and Supplementary Fig. S2). Also, the scalability of the DeepHINT model regarding the input sequence length and the number of training samples were evaluated (Supplementary Fig. S3). Moreover, a series of statistical analyses were performed to show the associations between the DeepHINT prediction scores and experimentally derived genomic features, further supporting the biological relevance of its prediction results (Supplementary Notes and Supplementary Fig. S4).

3.2 DeepHINT seq indicates important sequence positions for predicting HIV integration sites

The involvement of attention mechanism opened up the black box of deep learning and further enabled us to probe the derived attention map for each sample. Here we first try to test how DeepHINT model can learn the importance of features at individual positions with DNA sequence information as input alone. We hypothesize that the positions with larger attention values are more likely to associate with the sequence determinants of HIV integration site selection. Therefore, we were particularly interested in focusing on the *attention intensive regions* (which were defined as those the genomic positions possessing the highest 5% attention values of an input sequence) and examining how they can reflect the underlying biological mechanisms of HIV integration. Note that due to the convolution (whose kernel size was set to 6) and max pooling (whose pool size was set to 3) operations, each position in the attention map (also called the *attention map index*) represented an 8-bp region consisting of three continuous convolution kernels.

As a first attempt, we performed a close-up inspection for the distribution of the attention intensive regions near the integration sites for all positive and negative samples in the test dataset (Fig. 4a). Intriguingly, we observed a distinct pattern in the distributions of attention intensive regions between positive and negative samples. In particular, the distribution of attention intensive regions in positive samples showed a clear peak-valley-peak pattern near the integration sites, which, on the other hand, was not observed in negative samples. Such a discrepancy indicated that the attention map derived from our deep learning model was able to reflect the local sequence specificity pattern in the genomic context of HIV integration. Next, we further compared the above pattern derived from the attention intensive regions with the local consensus sequence motif obtained from the Score20 method, and found that the shape of the attention profile aligned well with the conserveness of each nucleotide in the Score20 motif (Fig. 4a and b). In particular, the sequence windows corresponding to attention map indexes 1 and 4, which included the most conservative G11 and C15 nucleotides in the Score20 consensus motif, showed the highest attention scores. On the other hand, the 'attention valley' in the observed attention profile also matched the non-conservative region in the Score20 motif (i.e. sequence windows corresponding to indexes 2 and 3). In addition, we also compared the distributions of attention values assigned to the Score20 regions (i.e. with attention map indexes 1 and 4) of integrations sites with either positive or negative Score20 values (Supplementary Fig. S5). Expectedly, the integration sites with a weaker Score20 motif (i.e. with negative Score20 values) tended to possess lower attention values ($P < 10^{-100}$ by two-sided Wilcoxon rank-sum test), indicating different molecular features within the genomic context and thus suggesting the possibility of being selected by other possible mechanisms. A representative example of integration sites possessing the attention intensive regions near the integration site can be found in Supplementary Figure S6.

Given the great improvement in the prediction performance of DeepHINT over the Score20 method (Fig. 3a and b), we were interested in how DeepHINT learns the sequence features beyond the Score20 region. Therefore, we further examined the associations of attention intensive regions with diverse genomic profiles, and showed that the important genomic positions indicated by the

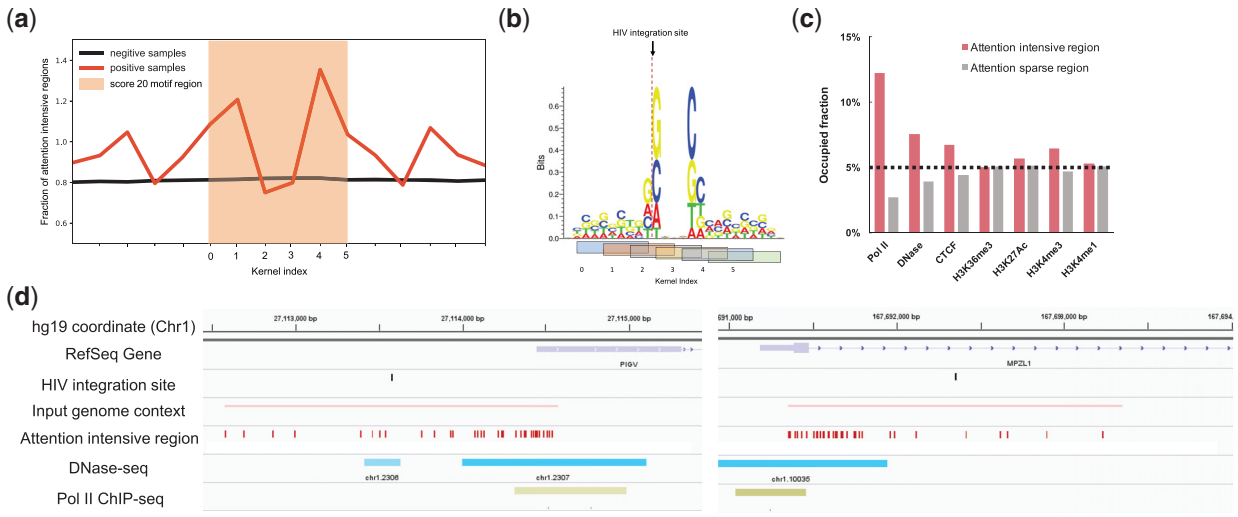

Fig. 4. Attention intensive regions indicate important local features of the predicted HIV integration sites. (**a**) A local view of the distribution of attention intensive regions near the integration sites for both positive samples (red) and the matched random control sites of negative samples (black) in the test dataset. The region overlapping with the Score20 consensus motif is highlighted (orange). Fractions of attention intensive regions were averaged among all the samples and normalized to the mean of all positions. Note that due to the convolution (with kernel size of 6) and max pooling (with pool size of 3) operations, each position in the attention map (termed as the attention map index) represented an 8-bp region consisting of three continuous convolution kernels. (**b**) The Score20 consensus motif after being aligned with the average attention profile. The attention window index corresponding to (a) is shown as x-axis and the sequence windows corresponding to each individual indexes are labeled below. The DNA WebLogo is visualized using Seq2Logo (Thomsen and Nielsen, 2012). (**c**) The fractions of different ChIP-seq peaks that are occupied by either attention intensive regions (with the highest 5% attention, pink) or attention sparse regions (with the lowest 5% attention, gray). (**d**) Two representative examples illustrating the enrichment of attention intensive regions in Pol II binding and DNase hypersensitive regions. The visualization was conducted using the IGV browser (Robinson *et al.*, 2011)

attention mechanism can provide more mechanistic insights into the integration process. In particular, we assessed how much of the genomic feature peaks (i.e. ChIP-seq peaks) within the 2000-bp context of an integration site are occupied by the attention intensive regions (Fig. 4c). Note that here we cannot conclude the importance of each feature to HIV integration site selection through this analysis because the comparisons were not between the integration sites and background sites and we only considered effect of each genomic factors within the peak regions. Here, we aimed to examine the associations between attention values and individual genomic features. In particular, we found that Pol II binding sites exhibited a significant association with the attention intensive regions, showing a 4.5-fold of enrichment compared to that of the attention sparse regions. In addition, the DNase hypersensitive regions also displayed a 1.9-fold of difference between the attention intensive and sparse regions, further supporting the potential ability of DeepHINT to capture the binding of regulatory factors in the chromatin accessible regions for HIV integration. In comparison, for the peaks of histone markers, i.e. H3K36me3, H3K27ac and H3K4me3, which commonly spread widely across the sequence context of HIV integration, we did not detect large signal difference between attention intensive and sparse regions (Fig. 4c), which also indicated the necessity to examine the more localized genomic features, e.g. sequence motifs, that may be more easily captured by the attention mechanism. Two representative examples of the enriched attention aligned with the Pol II binding and DNase hypersensitive regions can be found in Figure 4d.

3.3 DeepHINT highlights the sequence features for HIV integration site selection

To further exploit the specific sequence features captured by our attention mechanism, we also conducted a systematic survey on the sequence enrichment in those attention intensive regions. More specifically, we extracted all the 8-bp sequences in the attention intensive regions and calculated the enrichment of the binding motifs of known mammalian DNA binding proteins using HOMER (Heinz *et al.*, 2010). Importantly, as the attention values only indicate the importance of each positions within the input sequence context (i.e. they are non-negative values serving as the weights for combining all local features from different positions), they do not indicate whether a specific sequence motif is enriched or depleted near the integration sites when compared to the control sites. With reference to the control sites, we further calculated the odds ratio of being an integration site regarding the presence or absence each motif uncovered by DeepHINT and the corresponding *P* values derived from Chi-square tests to confirm the specific role of each uncovered sequence motif in HIV integration site selection (Table 1).

Intriguingly, we identified several important regulatory factors, whose binding sites showed significant enrichment in the attention intensive regions (Fig. 5a and b). In particular, we found in the

Table 1. Odds ratio of being integration sites with respect to the presence of sequence motifs uncovered by DeepHINT

Motif	Odds ratio	*P* value
THRb	1.52	$<1 \times 10^{-300}$
ZNF711	1.10	9.23×10^{-49}
BMAL1	1.06	7.58×10^{-9}
ZFX	1.03	2.69×10^{-4}
Maz	0.91	2.03×10^{-54}
Foxo3	0.91	7.48×10^{-42}
Tgif2	0.86	1.42×10^{-117}

Note: The presence of each sequence motif was calculated using FIMO (Grant *et al.*, 2011) with the default setting. *P* values are calculated by Chi-square test.

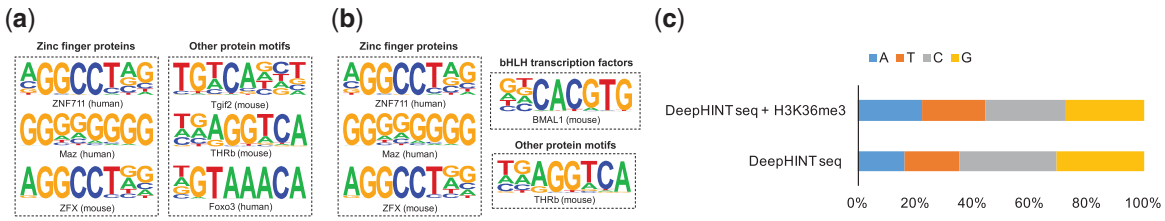

Fig. 5. Attention intensive regions are enriched with known regulatory motifs. The known mammalian motifs obtained from TRANSFAC (Matys *et al.*, 2006) that were significantly enriched in attention intensive regions are shown. All the motifs showed a Benjamini q-value $< 10^{-4}$ as determined by HOMER (Heinz *et al.*, 2010) and further evaluated by Chi-square to show significant enrichment or depletion near integration sites (P value $< 10^{-3}$). (**a**) The sequence motifs output by DeepHINT with DNA sequence alone as input. (**b**) The sequence motifs output by DeepHINT with DNA sequence plus H3K36me3 ChIP-seq profiles as input. (**c**) Comparison of nucleotide composition in the attention intensive regions using different input profiles

literature that zinc finger protein ZFX (Gazin, 1999) and thyroid hormone receptor (Hsia and Shi, 2002) have been reported to regulate the long terminal repeats (LTRs) mediated transcription. Similarly, the presence of E box motif, the core binding element of bHLH transcription factors, in LTRs has also been shown to be associated with the modulation of gene expression and maintenance of virus latency (Ou *et al.*, 1994), probably by repressing the expression of viral proteins (Jiang *et al.*, 2007; Terme *et al.*, 2009). However, how these binding elements present in the human genome sequence participate in the regulation of HIV latency still remains unclear. Despite the necessity of further experimental validation, our results may provide several possible directions to further explore cis-regulatory factors in human genome that may contribute to HIV integration site selection.

In addition, we were also interested in the difference of the attainable regions with different input profiles, i.e. DNA sequence alone or DNA sequence plus H3K36me3 profiles. Remarkably, although the sequence motif uncovered by the attention mechanism are largely overlapping with those generated by DNA sequence as input alone (Fig. 5a and b), we noticed that the most significant change compared with the sequence-only model was the increase of A/T rich regions (Fig. 5c). This observation was consistent with association between the H3K36me3 marks and LEDGF/p75 (Pradeepa *et al.*, 2012), the most important transcription factor for HIV binding which prefers to bind at A/T rich regions with its AT hook motif (Poeschla, 2008). Therefore, we reasoned that the introduction of H3K36me3 signals may push the attention mechanism to focus more on additional context regions, especially the A/T-rich regions, to further boost the prediction performance. These results also indicated the importance of incorporating various sources of information as well as introducing an attention layer to enhance the explainability of our deep learning model.

4 Discussion

Despite the long-lasting experimental and computational effort devoted to study HIV integration (Berry *et al.*, 2006; Brady *et al.*, 2009; Santoni *et al.*, 2010; Singh *et al.*, 2015), our current understanding of the mechanistic implications of HIV genome integration still remains largely limited. In this study, instead of focusing on the tedious handcrafted feature engineering, we developed an attention-based deep learning framework, namely DeepHINT, to automatically learn the contextual sequence features of HIV integration, and precisely predict the integration sites. In addition to boosting the prediction performance, the attention map derived by DeepHINT can explicitly decode how the deep learning model recognized highly relevant sequence features at different positions for final prediction, enabling one to better understand the underlying mechanism of the HIV integration.

Given the black box nature of deep learning, in this work, we tried to interpret the DeepHINT prediction results from two different aspects. First, through a subgroup analysis, we evaluated conditional effect of each genomic feature on the DeepHINT score, which can be regarded as a quantification of how likely a specific site tends to be an HIV integration site (Supplementary Fig. S4). This analysis was excepted to reflect how much some experimentally derived genomic features are associated with the DeepHINT score and thus enhance its biological relevance. Second, to embed the explainability in our framework, we introduced the attention mechanism, which has been widely used in the deep learning community to indicate important positions in the raw input. In fact, these two approaches are complementary to each other in generating biological insights from the deep learning framework. In particular, the first approach associates each genomic site with a specific score, which can be analyzed together with any given experimentally derived feature. On the other hand, by uncovering sequence motifs enriched in the attention intensive regions, we can identify unexpected features that may also play an important role in HIV integration site selection. However, there are also limitations in the attention mechanism. In particular, each attention value corresponds to a specific window in the input sequence context, which can only be interpreted as a localized quantification of the importance. More importantly, since the attention values only indicate the importance of each position within the input context, further efforts are still needed to calculate the statistical enrichment or depletion of each sequence motif uncovered by the attention mechanism to confirm its specific role.

The current study demonstrates a usage of deep learning, especially with the attention mechanism, for predicting and analyzing HIV integration sites. Admittedly, the further explorations of the underlying mechanisms of HIV integration would rely on the generation of more high-quality HIV integration sites, especially in human patients (Maldarelli *et al.*, 2014; Wagner *et al.*, 2014). Considering the model complexity of deep neural networks, it is always necessary to collect a large amount of training samples to fully exploit their prediction power, which in fact can be reflected by the decreased performance when a limited number of training samples were used (Supplementary Fig. S3). In addition, the introduction of effective machine learning techniques, e.g. using transfer learning to transfer the cell line knowledge to patient samples, will also be an interesting future direction to pursue. As integration-associated virus latency is attracting more and more research interest (Demeulemeester *et al.*, 2015), we believe that our DeepHINT framework together with more emerging experimental data

(Shao *et al.*, 2016) and improved experimental techniques (Sherman *et al.*, 2017) will offer more useful insights into the studies of HIV integration in the genome.

Funding

This work was supported in part by the National Natural Science Foundation of China (61472205, 81630103, 61772197 and 81530065), China's Youth 1000-Talent Program, Beijing Advanced Innovation Center for Structural Biology and US National Science Foundation IIS-1646333. We acknowledge the support of NVIDIA Corporation with the donation of the Titan X GPU used for this research.

Conflict of Interest: none declared.

References

Alipanahi,B. *et al.* (2015) Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nat. Biotechnol.*, **33**, 831–838.

Bahdanau,D. *et al.* (2014) Neural machine translation by jointly learning to align and translate. In: *Conference Paper at International Conference on Learning Representations (ICLR) 2015*.

Bengio,Y. (2012) Neural Networks: Tricks of the Trade. In: *Practical Recommendations for Gradient-Based Training of Deep Architectures*. 2nd edn, Springer, Berlin, Heidelberg, pp. 437–478.

Berry,C. *et al.* (2006) Selection of target sites for mobile DNA integration in the human genome. *PLoS Comput. Biol.*, **2**, e157.

Brady,T. *et al.* (2009) HIV integration site distributions in resting and activated CD4+ T cells infected in culture. *AIDS (London, England)*, **23**, 1461.

Ciuffi,A. *et al.* (2005) A role for ledgf/p75 in targeting HIV DNA integration. *Nat. Med.*, **11**, 1287–1289.

Demeulemeester,J. *et al.* (2015) Retroviral integration: site matters. *Bioessays*, **37**, 1202–1214.

Deming,L. *et al.* (2016) Genetic architect: discovering genomic structure with learned neural architectures. arXiv preprint arXiv, 1605.07156.

Gazin,C. (1999) ZFX transactivation of the HIV-1 LTR is cell specific and depends on core enhancer and TATA box sequences. *Nucleic Acids Res.*, **27**, 2156–2164.

Grant,C.E. *et al.* (2011) Fimo: scanning for occurrences of a given motif. *Bioinformatics*, **27**, 1017–1018.

Heinz,S. *et al.* (2010) Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and b cell identities. *Mol. Cell*, **38**, 576–589.

Hsia,S.-C.V. and Shi,Y.-B. (2002) Chromatin disruption and histone acetylation in regulation of the human immunodeficiency virus type 1 long terminal repeat by thyroid hormone receptor. *Mol. Cell Biol.*, **22**, 4043–4052.

Jiang,G. *et al.* (2007) c-MYc and Sp1 contribute to proviral latency by recruiting histone deacetylase 1 to the human immunodeficiency virus type 1 promoter. *J. Virol.*, **81**, 10914–10923.

Lusic,M. and Siliciano,R.F. (2017) Nuclear landscape of HIV-1 infection and integration. *Nat. Rev. Microbiol.*, **15**, 69–82.

Maldarelli,F. *et al.* (2014) Specific HIV integration sites are linked to clonal expansion and persistence of infected cells. *Science*, **345**, 179–183.

Mao,W. *et al.* (2017) Modeling enhancer-promoter interactions with attention-based neural networks. bioRxiv, 219667.

Marini,B. *et al.* (2015) Nuclear architecture dictates HIV-1 integration site selection. *Nature*, **521**, 227–231.

Matys,V. *et al.* (2006) Transfac® and its module transcompel®: transcriptional gene regulation in eukaryotes. *Nucleic Acids Res.*, **34**, D108–D110.

Ou,S. *et al.* (1994) Role of flanking e box motifs in human immunodeficiency virus type 1 tata element function. *J. Virol.*, **68**, 7188–7199.

Pan,X. and Yan,J. (2017) Attention based convolutional neural network for predicting RNA-protein binding sites. arXiv preprint arXiv, 1712.02270.

Poeschla,E.M. (2008) Integrase, ledgf/p75 and hiv replication. *Cell. Mol. Life Sci.*, **65**, 1403–1424.

Pradeepa,M.M. *et al.* (2012) Psip1/Ledgf p52 binds methylated histone H3K36 and splicing factors and contributes to the regulation of alternative splicing. *PLoS Genet.*, **8**, e1002717.

Quang,D. and Xie,X. (2016) DanQ: a hybrid convolutional and recurrent deep neural network for quantifying the function of DNA sequences. *Nucleic Acids Res.*, **44**, e107.

Robinson,J.T. *et al.* (2011) Integrative genomics viewer. *Nat. Biotechnol.*, **29**, 24.

Rumelhart,D.E. *et al.* (1986) Learning representations by back-propagating errors. *Nature*, **323**, 533–536.

Santoni,F.A. (2013) EMdeCODE: a novel algorithm capable of reading words of epigenetic code to predict enhancers and retroviral integration sites and to identify H3R2me1 as a distinctive mark of coding versus non-coding genes. *Nucleic Acids Res.*, **41**, e48.

Santoni,F.A. *et al.* (2010) Deciphering the code for retroviral integration target site selection. *PLoS Comput. Biol.*, **6**, e1001008.

Schröder,A.R. *et al.* (2002) HIV-1 integration in the human genome favors active genes and local hotspots. *Cell*, **110**, 521–529.

Serrao,E. *et al.* (2014) Integrase residues that determine nucleotide preferences at sites of HIV-1 integration: implications for the mechanism of target DNA binding. *Nucleic Acids Res.*, **42**, 5164–5176.

Shao,W. *et al.* (2016) Retrovirus integration database (rid): a public database for retroviral insertion sites into host genomes. *Retrovirology*, **13**, 47.

Sherman,E. *et al.* (2017) INSPIIRED: a pipeline for quantitative analysis of sites of new DNA integration in cellular genomes. *Mol. Ther. Methods Clin. Dev.*, **4**, 39–49.

Singh,P.K. *et al.* (2015) LEDGF/p75 interacts with mRNA splicing factors and targets HIV-1 integration to highly spliced genes. *Genes Dev.*, **29**, 2287–2297.

Singh,R. *et al.* (2017) Attend and predict: understanding gene regulation by selective attention on chromatin. In: *Advances in Neural Information Processing Systems*, Curran Associates, pp. 6788–6798.

Srebro,N. *et al.* (2005) Maximum-margin matrix factorization. *Adv. Neural Inform. Process. Syst.*, **17**, 1329–1336.

Srivastava,N. *et al.* (2014) Dropout: a simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.*, **15**, 1929–1958.

Terme,J.-M. *et al.* (2009) E box motifs as mediators of proviral latency of human retroviruses. *Retrovirology*, **6**, 81.

Thomsen,M.C.F. and Nielsen,M. (2012) Seq2Logo: a method for construction and visualization of amino acid binding motifs and sequence profiles including sequence weighting, pseudo counts and two-sided representation of amino acid enrichment and depletion. *Nucleic Acids Res.*, **40**, W281–W287.

Vijaya,S. *et al.* (1986) Acceptor sites for retroviral integrations map near DNase I-hypersensitive sites in chromatin. *J. Virol.*, **60**, 683–692.

Wagner,T.A. *et al.* (2014) Proliferation of cells with HIV integrated into cancer genes contributes to persistent infection. *Science*, **345**, 570–573.

Wallace,B. *et al.* (2011) Class imbalance, redux. In: *2011 IEEE 11th International Conference on Data Mining*, pp. 754–763.

Wang,G.P. *et al.* (2007) HIV integration site selection: analysis by massively parallel pyrosequencing reveals association with epigenetic modifications. *Genome Res.*, **17**, 1186–1194.

Wong,J.K. *et al.* (1997) Recovery of replication-competent HIV despite prolonged suppression of plasma viremia. *Science*, **278**, 1291–1295.

Zhang,S. *et al.* (2015) A deep learning framework for modeling structural features of RNA-binding protein targets. *Nucleic Acids Res.*, **44**, e32–e32.

Zhang,S. *et al.* (2017a) Analysis of ribosome stalling and translation elongation dynamics by deep learning. *Cell Syst.*, **5**, 212–220.

Zhang,S. *et al.* (2017b) TITER: predicting translation initiation sites by deep learning. *Bioinformatics*, **33**, i234–i242.

Zhou,J. and Troyanskaya,O.G. (2015) Predicting effects of noncoding variants with deep learning-based sequence model. *Nat. Methods*, **12**, 931–934.